Homework 3

EAEEE 4257 Environmental Data Modeling and Analysis

Due Wednesday, February 28, 2018

This homework will give you more practice with regression modeling in a more open-ended context. These two problems will give you a chance to explore real-world data, explore the data, build a model, check your model, and change your model. There is not a single "right" way to answer these questions. Instead, look at the data, show your work, use the ideas we have discussed in class, state your assumptions, check your assumptions against the data, and come up with something that you think is reasonable.

Please show your work: if you build 11 models and perform a sophisticated analysis to choose the best one, but your final writeup doesn't include these models, your grader won't know that you built them! Please also remember to include text to describe the modeling choices you are making, why you are making them, and why you think they are justified. A good thought process is more important than the R^2 of your final model!

Please work in `R Markdown` as you have done so far this semester. You will turn in a separate file for each problem; you can use any of the `R Markdown` files we have used this semester (such as the HW2 file) as a template or you can use another template, as long as it knits to `html`. Please also make sure that you download the raw data files to the same folder as your `.Rmd` files so that you don't have long paths in your `readr::read_csv` command (this lets the grader re-run your code without changing anything). Please also look at the `read_data.Rmd` and `regression_tools.Rmd` files for some helpful codes to read in the raw data and build linear regressions. If you would like more codes for linear regression, post in the `R-Computing` Slack channel and it will be added to `regression_tools.Rmd` on Courseworks.

1 Aquifer Pollutants

The Ogallala aquifer was investigated to determine relationships between uranium and other concentrations in its waters. You can find this data in `appc16.csv`. Construct a

regression model to predict uranium as a function of total dissolved solids and the presence of bicarbonate.

2 Aquifer Pollutants

The file `appc15.csv` contains data from 42 small urban drainage basins located in several cities around the United States (Mustard et al., 1987). The variable we are interested in is called `nitrogen` and is the total nitrogen load for the basin. There are eight possible explanatory variables to use for prediction purposes, summarized in `read_data.Rmd`. Pay special attention to multi-collinearity to come up with a model which allows you to predict nitrogen load given the other predictors in the data set.

A Note

There is some discussion of these data sets in the Helsel and Hirsch book. You can read their analysis and consider it for your own analysis, but their approach is not the "only" (or probably even the "best") way to approach this problem. Again, make sure you include substantial *text* in your R markdown document describing your thought process!

Summary

The things you need to turn in as part of this assignment are:

1. `hw3_prob1.Rmd`: your analysis as a `R Markdown` for problem 1
2. `hw3_prob1.html`: the knitted version of your analysis for problem 1
3. `hw3_prob2.Rmd`: your analysis as a `R Markdown` for problem 2
4. `hw3_prob2.html`: the knitted version of your analysis for problem 2

Please upload these as separate files to courseworks: *zipped folders will not be accepted!* Make sure that you include your name and UNI on the document.

If you have difficulty with this homework, particularly with using the Data Camp website or installing `R` and the required packages, please use the `r-computing` channel on the Slack page.